一种基于语义组块特征的改进 Cosine 文本相似度计算方法*

白如江[1]　冷伏海[2]　廖君华[1]
[1](山东理工大学科技信息研究所　淄博　255049)
[2](中国科学院科技战略咨询研究院　北京　100190)

摘要：【**目的**】利用文本语义组块特征提升 Cosine 文本相似度计算性能。【**方法**】获取 NSF 资助的关于碳纳米管研究领域的项目数据，进行词干还原、词性标注等预处理；利用条件随机场模型实现文本内容的语义组块标注；在此基础上实现基于语义组块特征的改进 Cosine 文本相似度计算，并与未标注的数据进行相似度计算比较，分析实验结果。【**结果**】实验证明基于语义组块特征的改进 Cosine 相似度计算结果比原始文本 Cosine 相似度计算结果相似度均有不同程度的提升，在实验数据中最高的相似度提升了 26%。【**局限**】依赖于语义组块标注性能。【**结论**】本文方法能有效提升文本间语义相似度，降低向量空间模型维度，提高计算效率，并且具有良好的泛化能力和鲁棒性。

关键词：文本相似度　语义组块　向量空间模型　本体
分类号：G250

1　引　言

相似性是自然世界中普遍存在的一种关系，在现实世界中任意两个对象之间或多或少存在一定相似性关系。相似性的大小可以用相似度定量表示。自由文本之间同样存在复杂的相似性关系，在自然语言处理中，需要把这种复杂的关系用一种简单的数量来度量，文本相似度计算应运而生。文本相似度计算是数据挖掘、人工智能、信息检索等领域研究的基本问题。随着文本相似度研究的不断深入，相似度计算的对象由词共现相似、语法结构相似上升到语义相似。而精准高效的语义相似度计算成为一个亟待解决的问题。

在科学研究前沿探测、研究主题演化、主题聚类识别等情报学研究领域，文本相似度计算也无处不在。具体来讲，通过文本相似度计算可以发现：不同研究主题之间的相互关联影响情况；不同时间段内同一研究主题演化发展变化情况；学科交叉研究主题重合情况；通过与现有研究主题相似度对比发现新兴研究主题；不同文献研究内容相似度情况等一系列问题。

因此，一种富含语义关联度的高效文本相似度计算方法可以有效地发现不同文本之间的相互关系，在情报学研究领域，可以利用此方法充分挖掘文本内容关系，从而可以帮助提升科技情报研究的准确性和前瞻性。目前，基于向量空间模型相似度计算方法存在语义信息缺失，向量维度过高，基于本体语义相似度计算方法又存在过分依赖外在本体的问题。因此，本文尝试利用语义组块标注信息改进 Cosine 相似度计算方法，以期提升文本相似度计算性能。

2　相关研究

在文本相似度计算领域主要计算方法归纳起来可

通讯作者：白如江，ORCID: 0000-0003-3822-8484，E-mail: brj@sdut.edu.cn。
*本文系国家社会科学基金项目“未来新兴科学研究前沿识别研究”(项目编号：16BTQ083)的研究成果之一。

以分为三种: 基于几何向量空间的计算方法; 基于词项统计的计算方法; 基于本体的计算方法。

2.1　基于几何向量空间的方法

基于几何向量空间的文本相似度计算方法的基础是将需要计算相似度的文档以向量空间模型(Vector Space Model, VSM)表示。向量空间模型是由信息检索领域著名专家 Salton 等提出, 并成功地应用于著名的 SMART 文本检索系统[1]。

在向量空间模型中, 文档被映射到由词项构成的几何空间中。向量空间模型可以用公式(1)表示。

$$D = D(T_1, W_1; T_2, W_2; \cdots; T_n, W_n) \qquad (1)$$

其中, D(*Document*)表示文档; T(*Term*)表示文档中的词项; W(*Weight*)表示词项在文档中的权重; n 为文档中的词项数量。

利用几何向量空间计算相似度的模型主要有: 欧几里得距离(Euclidean Distance)和余弦相似度(Cosine Similarity)[2]。

2.2　基于词项统计的方法

与基于几何向量空间模型计算方法不同, 基于词项统计的计算方法主要考虑词项在文本中所占的比例, 如果两个文档中所共同包含的词项越多, 那么这两个文档就越相似。基于此项统计的方法主要有两种思路。

(1) 基于重叠词的方法

该方法认为两段文本所构成的词或短语重叠个数越多则两段文本的相似度就越大。这种方法的最具代表性算法是 Jaccard 相似度系数方法[3]。

此外, 基于重叠词思想上实现的算法还有简单词重叠法[4]、IDF 重叠法[5]以及 Zipfian 重叠法[6-7]。

(2) TF-IDF 及其各种加权算法

TF-IDF (Term Frequency-Inverse Document Frequency)是一种统计方法, 用以评估一个字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加, 但同时会随着它在语料库中出现的频率成反比下降。

利用 TF-IDF 各种形式加权的代表方法有 LSA[8]、HAL[9]、Islam 等的方法[10]以及 Allan 等提出的方法[7]等。

2.3　基于本体的方法

基于本体的文本语义相似度计算方法主要是利用本体库或语义词典内蕴含的丰富的语义信息来提高文本语义相似度。该方法可以归纳为三种。

(1) 基于本体库边距离的计算方法

大多本体库或语义词典(如WordNet①)将相关词和概念词组织在一棵或几棵树状的层次结构中。在一棵树状图中, 任何两个节点之间可以通过一条路径连接, 并且这条路径是唯一的。基于本体库边深度的计算方法认为, 这条路径的长度可以衡量这两个节点(词语、概念)间语义距离。随着概念所在的节点距离根节点越深, 其所包含的语义信息越丰富。代表算法[11]有: Rada等[12]、Leacock等[13]、Pekar等[14]。Rada等认为衡量两个概念词间的相似度可以通过计算其在本体分类体系树中的最短距离获得[12]。后来对Rada距离的改进主要集中在Rada所有边的距离同等重要的假设改进[15]。

(2) 基于本体库节点的计算方法

与基于边计算相似度思路不同, 基于节点的相似度主要考虑文本中的词或词组在本体库中概念的对应关系, 从而计算语义相似度。基于本体库节点的计算方法根据对节点的计算方法不同可以分为基于节点特征策略和基于信息熵的策略。基于节点特征策略思想来源于 Tversky[16]提出的特征模型(Feature-model)。在本体库中主要是考虑当前节点的父亲节点以及往上的祖先节点和根节点。

(3) 混合计算方法

混合方法是将本体库中基于边的计算策略和基于节点的计算策略共同考虑建立起来的一种方法。通常会通过对边、节点的权重调节计算文本的相似度[17-19]。

由于语言的差异, 中文文本相似度计算研究主要是根据中文特点提出相关相似度计算模型[20-23]。

综上所述, 在相似度计算方法上, 基于几何向量空间的计算方法和基于词项统计的计算方法都忽略了词项本身的语义信息。在计算过程中主要依据词项是否都会出现在两段文本中, 而且要求词项必须完全相同。由于同一个概念可能会以不同的表达方式出现(如“计算机”可以表述为“电脑”、“笔记本”等)或同

①http://wordnet.princeton.edu/.

一词项在不同的上下文有不同的语义解释(如“苹果”可以是一种水果，也有可能是手机)，这种情况下会严重影响相似度计算的准确性。此外，基于向量空间模型的方法在处理长文档的时候，由于文档过长，生成的向量空间维度过高，在计算文本相似度的时候可以度量的词项过少，从而造成文档相似度计算不够准确。基于本体的计算方法在一定程度上克服了上述两种方法的缺点，但是需要依赖外在本体库或语义词典。

因此，为了克服向量空间模型语义信息缺失、向量维度过高以及基于本体语义相似度计算方法依赖外在本体的问题，本文提出一种基于语义组块特征的改进 Cosine 文本相似度计算方法。该方法与向量空间模型相比，能够在一定程度上反映文档的语义信息，并且可以通过语义向量空间分割有效降低向量空间模型维度。与基于本体的计算方法相比，该方法不需要外部的本体库支持，模型的泛化能力得到提高。

3 基于语义组块特征的改进 Cosine 文本相似度计算

3.1 语义组块特征分析

文本中经常存在这样的情况，即文本论述的内容词汇大部分相同，只有部分词汇不同。但是正是这些不同的词汇具有极强的语义功能。如 S1 和 S2 两个句子。

S1: The main goal of this NSF project is to develop new class of hybrid composite structures.

S2: The main goal of this NSF project is to develop new class of singlewalled carbon nanotube.

两句中前 13 个词汇完全一致，只是后 3 个词汇不同，正是这后面 3 个词汇说明了该项目的研究目的有着根本的不同，本文将这 3 个词汇定义为“研究目的”语义组块，如图 1 所示。

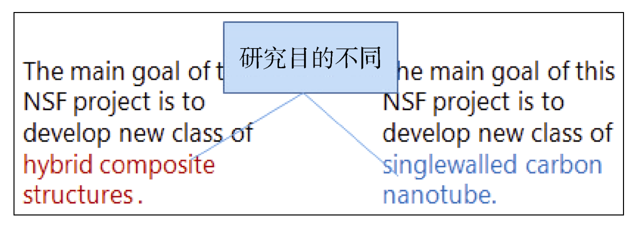


图 1　不同研究目的的语义组块

在科技文本数据中还存在另外一种情况，文本论述的词汇字面相同，但是所起到的语义角色不同，如 S3 和 S4 两个句子。

S3: The main goal of this NSF project is to improve Chemical Vapor Deposition method.

S4: The main goal of this NSF project is apply Chemical Vapor Deposition method to develop SWCNT.

两个句子都提到了“Chemical Vapor Deposition method”这种方法，但是在 S3 中该组块的语义角色是“研究目的”，在 S4 中该组块的语义角色是一种“研究方法”，研究如何利用该方法制备“单壁碳纳米管，SWCNT”，如图 2 所示。

The main goal of t 研究目的 NSF project is to improve Chemical Vapor Deposition method.
The 研究方法 of this NSF pro... Chemical Vapor Deposition method to develop SWCNT

图 2　具有不同语义角色的语义组块

3.2 改进 Cosine 相似度计算方法

为了利用上述语义组块特征提升文本相似度性能，本文提出一种基于语义组块特征的文本相似度计算方法。

假设 A,B 两个项目文档，A 文档中有 n 个词语，B 文档中有 m 个词语，如公式(2)所示。

$$
\begin{aligned}
A &= \{w_{11}, w_{12}, \cdots w_{1n}\} \\
B &= \{w_{21}, w_{22}, \cdots w_{2m}\}
\end{aligned} \tag{2}
$$

其中，A,B 分别表示两个项目文档；$w_{11},w_{12},\cdots,w_{1n}$ 表示文档 A 中的所有词语；$w_{21},w_{22},\cdots,w_{2m}$ 表示文档 B 中的所有词语。

根据文本中的不同语义组块角色信息，比如可以文本中蕴含的“研究目的”、“研究方法”、“实验设备”、“实验材料”等语义组块信息，将上述空间划分为不同语义角色的文本向量表示，如公式(3)所示。

$$
\begin{aligned}
A &= \{w_{11}, w_{12}, \cdots w_{1n}\}SC1 \cup \{w_{11}, w_{12}, \cdots w_{1n}\}SC2 \cdots \\
&\quad \cup \{w_{11}, w_{12}, \cdots w_{1n}\}SCn \\
B &= \{w_{21}, w_{22}, \cdots w_{2m}\}SC1 \cup \{w_{21}, w_{22}, \cdots w_{2m}\}SC2 \cdots \\
&\quad \cup \{w_{21}, w_{22}, \cdots w_{2m}\}SCn
\end{aligned} \tag{3}
$$

其中，A,B 分别表示两个不同的科技文档；$\{w_{11},w_{12},\cdots,w_{1n}\}_{SC1}$ 表示文档 A 中属于 $SC1$ 语义角色特征的词汇集合；$\{w_{11},w_{12},\cdots,w_{1n}\}_{SCn}$ 表示文档 A 中属于

SCn 语义角色特征的词汇集合。以此类推。

由于Cosine距离计算函数在文本相似度计算方面表现出的良好性能[2]，本文利用 Cosine 相似度计算函数，结合前面标注出的语义组块特征，将向量空间进行了“研究目标”、“研究方法”等语义功能分割，并在此基础上实现项目文档数据语义相似度计算模型，如公式(4)所示。

$$Sim_semantic(A,B)=\sum_{j=1}^{k}\left(\frac{\sum_{i=1}^{n}(A_i\times B_i)}{\sqrt{\sum_{i=1}^{n}A_i^2\times\sum_{i=1}^{n}B_i^2}}\right) \quad (4)$$

其中，*Sim_semantic*(*A*，*B*)表示两个项目文档 *A*,*B* 之间的语义相似度；*j* 为语义组块特征，$j\in\{SC1, SC2,\cdots,SCn\}$；*n* 为文档中词项的数量；$A_i$表示项目文档 *A* 中的第 *i* 个词项；$B_i$表示项目文档 *B* 中的第 *i* 个词项。

4　实　验

4.1　数据集

本文从美国自然科学基金网站①上在项目标题和摘要中出现“Carbon Nanotube”或“CNT”关键字为检索策略进行检索，总共得到 615 条数据，并对其进行语义组块标注[24]。每个项目数据大概包含 500 个单词。为了验证本文提出的方法是否能够有效提升文本之间语义相似度计算效能，从中随机选取6个项目数据进行人工判读，再利用本文提出的方法进行实验，验证本文提出方法的有效性。这 6 个项目的基本信息如表 1 所示。

表 1　NSF 项目基本信息

AwardNumber	Title	Program(s)
0933141	Novel Catalyst Supports for Water Electrolysis: Experimental and Theoretical Studies	ENERGY FOR SUSTAINABILITY
0945004	SBIR Phase I: Low Density Carbon Fibers Based on Gel Spun Polyacrylonitrile/ Carbon Nanotube	SMALL BUSINESS PHASE I
1007793	Materials World Network: Novel Catalyst Systems for Carbon Nanotube (CNT) Synthesis and their Underlying Mechanisms	SOLID STATE & MATERIALS CHEMIS, OFFICE OF SPECIAL PROGRAMS-DMR
1046519	SBIR Phase I: Manufacturing of Double-Walled Carbon Nanotube/Rigid Rod Polymer Advanced Structural Fibers	SMALL BUSINESS PHASE I
1133117	Collaborative Research: Experimental and Theoretical Investigations of Catalysis on Carbon Nanotube Surfaces For Selective Liquid Fuel Generation	CATALYSIS AND BIOCATALYSIS
1434824	DMREF: Engineering Strong, Highly Conductive Nanotube Fibers Via Fusion	DMREF

4.2　实验平台

硬件环境：CPU：Intel®Core^{TM}i5-3317U 1.70GHz；内存：4.00GB；操作系统：Windows7 旗舰版 64 位；软件环境：Python4.3。

4.3　实验过程

(1) 数据预处理

利用 Knime 开源工具对文本进行词干还原、词性标注以及停用词去处等预处理，如图3所示。

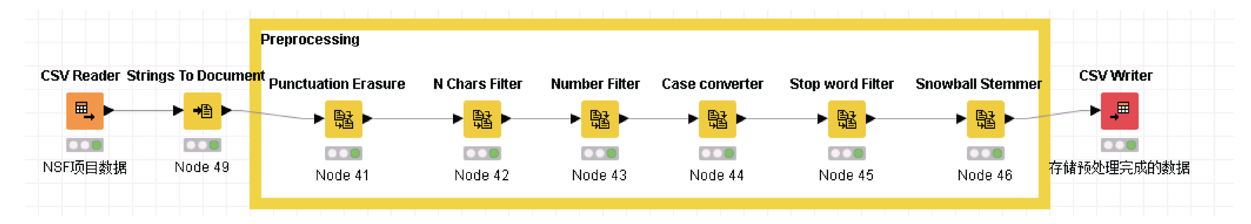


图 3　数据预处理

①https://www.nsf.gov/.

预处理完成后将所有数据存储到CSV文件中供下一步实验使用。

(2) 语义组块标注

利用 CRFSuite 开源工具[①]，基于条件随机场模型对实验数据进行语义组块标注，标注出“研究目的”、“研究方法”、“应用前景”、“研究性能”、“实验设备”、“实验材料” 6 种语义组块类型，为后续研究提供数据集支持，标注好的数据如图 4 所示。与语义组块标注相关详细内容请参考文献[24]。

9890	B-SUB	this	DT	B-NP	N1017000N_15
9891	I-SUB	project	NN	I-NP	N1017000N_15
9892	B-ACT	will	MD	B-VP	N1017000N_15
9893	I-ACT	develop	VB	I-VP	N1017000N_15
9894	B-GOL	a	DT	B-NP	N1017000N_15
9895	I-GOL	comprehensive	JJ	I-NP	N1017000N_15
9896	I-GOL	full-system	NN	I-NP	N1017000N_15
9897	I-GOL	simulation	NN	I-NP	N1017000N_15
9898	I-GOL	infrastructure	NN	I-NP	N1017000N_15
9899	I-GOL	that	WDT	B-NP	N1017000N_15
9900	I-GOL	consists	VBZ	B-VP	N1017000N_15
9901	I-GOL	of	IN	B-PP	N1017000N_15
9902	I-GOL	PCM	NNP	B-NP	N1017000N_15

图 4　语义组块标注结果(部分)

(3) 基于语义组块的改进 Cosine 相似度计算

通过 Python 平台，利用 sklearn 开源工具包[②]中的 cosine_similarity 模块实现基于语义组块的改进 Cosine 相似度计算，核心代码如下。

```
from sklearn.metrics.pairwise import cosine_similarity
data_path = r'E:\ SIM_EXPERIMENT\sim_txt_semantic_1'
filelist=os.listdir(data_path)
filenames=[os.path.join(data_path,f) for f in filelist]
vectorizer = CountVectorizer(input='filename')
dtm = vectorizer.fit_transform(filenames)
vocab=vectorizer.get_feature_names ()
vocab=np.array (vocab)
dist_cos = cosine_similarity(dtm)
```

(4) 相似度计算结果可视化展示

为了直观展示出不同文本之间采用不同相似度计算相似度结果变化情况，利用 matpolotlib[③]工具包对相似度计算结果在三维层面和二维层面进行可视化展示。

4.4　结果分析

经过上述实验步骤得到实验结果，如表 2 所示。第一列和第二列为项目文档编号，第三列(Raw_sim)为原始项目文本余弦相似度计算结果，第四列(Sem_sim)为本文提出的基于语义组块特征的项目文本 Cosine 相似度计算结果，第五列(Increase)给出了本文提出的基于语义组块特征的项目文本余弦相似度计算结果比原始项目文本余弦文本相似度计算结果的提升情况。

表 2　相似度计算实验结果

Doc_id	Doc_id	Raw_sim	Sem_sim	Increase
‘0933141’	‘0945004’	0.39	0.51	12%
‘0933141’	‘1007793’	0.63	0.71	8%
‘0933141’	‘1046519’	0.68	0.73	5%
‘0933141’	‘1133117’	0.42	0.69	**27%**
‘0933141’	‘1434824’	0.46	0.68	22%
‘0945004’	‘1007793’	0.4	0.51	11%
‘0945004’	‘1046519’	0.52	0.63	11%
‘0945004’	‘1133117’	0.26	0.51	**25%**
‘0945004’	‘1434824’	0.46	0.58	12%
‘1007793’	‘1046519’	0.63	0.74	11%
‘1007793’	‘1133117’	0.49	0.74	**25%**
‘1007793’	‘1434824’	0.52	0.68	16%
‘1046519’	‘1133117’	0.42	0.67	**25%**
‘1046519’	‘1434824’	0.57	0.73	16%
‘1133117’	‘1434824’	0.4	0.66	**26%**

表 3 为 6 个项目数据主要研究内容的人工判读结果。通过实验结果发现本文提出的基于语义组块特征的项目文本相似度比原始文本相似度计算性能有一定程度的提高。其中，‘1133117’项目与其他 5 个项目的相似度比例提升最多。分析其原因‘1133117’项目主要研究的是关于增强碳纤维(Carbon Fiber)/碳纳米管(Carbon NanoTube, CNT)分散颗粒的催化效率的问题。在这个研究项目中既涉及到碳纤维(Carbon Fiber)/碳纳米管(Carbon NanoTube, CNT)又涉及到催化剂(Catalysts)的问题，所以该项目与其他 5 个项目均有一定程度的联系。

①http://www.chokkan.org/software/crfsuite/.

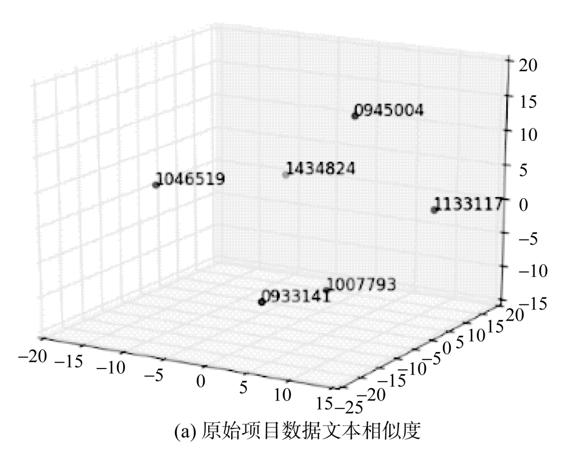
②http://scikit-learn.org/stable/.

③https://matplotlib.org/.

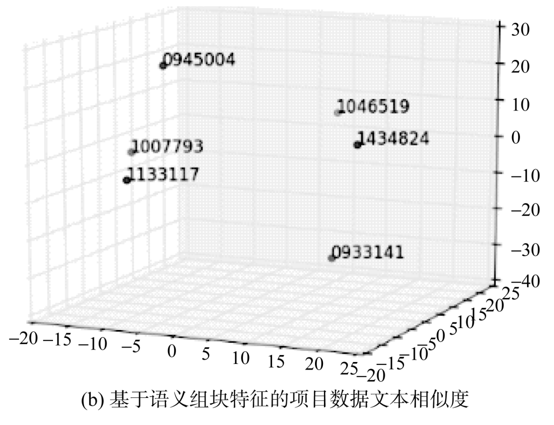
表 3 人工判读结果

项目编号	项目主要研究内容
0933141	开发一种新的纳米晶(Nano-Crystalline)混合金属氧化物催化剂(Oxide Catalyst)，能够获得理想的导电性和电化学特性。该项目还能够帮助理解在电化学过程中纳米材料结构对电化学稳定性和活性的影响。
0945004	利用凝胶纺丝技术(Gel Spun Technology)开发一种高强度–低密度的碳纳米管(Carbon NanoTube, CNT)基碳纤维(Carbon Fiber)。该纤维的拉伸强度大于 7Gpa，拉伸模量大于 450Gpa，密度小于 1.2g/cm^3。该纤维可以广泛应用于卫星、飞机机身、机翼以及高性能汽车中。
1007793	将探寻石墨烯(Graphene)和碳纳米管(Carbon NanoTube, CNT)对氧化物催化剂(Oxide Catalyst)影响机制，并关注新的在氧化物催化剂(Oxide Catalyst)新的增长变量。
1046519	利用高度结晶(Crystaline)的双壁碳纳米管(Double Wall Carbon Nanotube, DWCNT)制备具备高强度和韧性的新一代结构纤维。该纤维可以为车辆防弹、商业航空航天等领域提供强度更高、重量更轻的结构纤维材料。
1133117	研究碳纳米管(Carbon NanoTube, CNT)本身做为非均相催化反应(Heterogeneous Catalysis)尤其是在 FT 催化反应中的催化剂(Catalysts)的作用。增强碳纤维(Carbon Fiber)/碳纳米管(Carbon NanoTube, CNT)分散颗粒的催化效率。
1434824	一种新型碳纳米结构(Carbon Nanostructure)工程过程，称为纳米管融合(NanoTube Fusion)。该方法可以创建高性能的碳纤维(Carbon Fiber),可以应用于航空航天、高功率密度的能量存储和轻质布线等领域。

为了直观地表达各项目文档之间的相似度关系，本文在三维空间上标出了各项目文档所处的位置，如图 5 所示。每个点代表每个项目文档，文档之间的距离代表的文档之间的相似度，如果两个文档距离越近，那么说明这两个文档就越相似。



(a) 原始项目数据文本相似度



(b) 基于语义组块特征的项目数据文本相似度

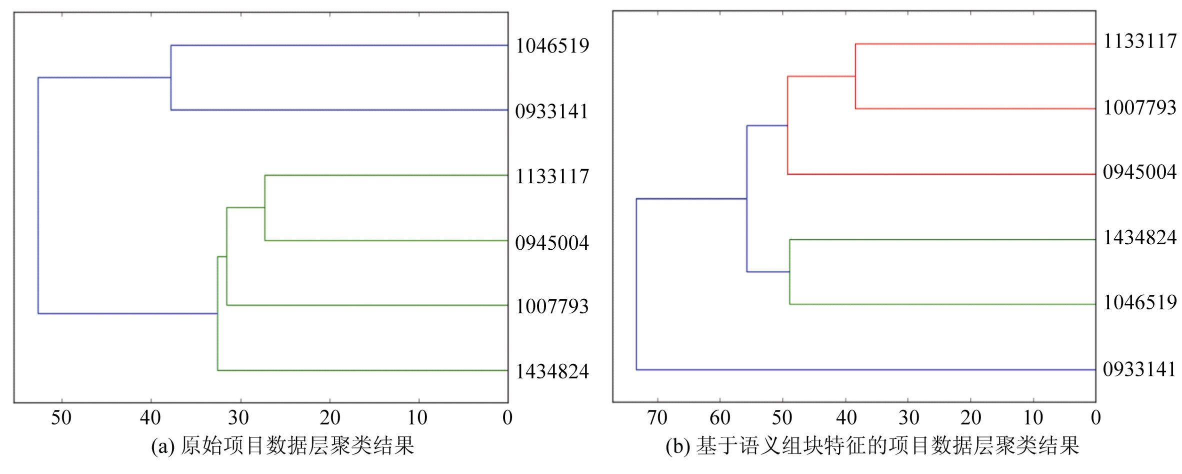
图 5 语义相似度距离对比

图5(a)是原始项目文本之间的相似度距离，图5(b)为经过语义组块标注后的相似度距离。可以看出，图5(a)中各个项目比较分散，很难看出不同文档之间的主题相似度关系。图5(b)中可以明显发现各个项目文档都向中心靠拢，各个项目之间的距离也相应缩小。变化最大的是'1133117'项目，明显地向'1007793'项目靠近。通过分析发现'1007793'项目研究的是探寻石墨烯(Graphene)和碳纳米管(Carbon NanoTube, CNT)对氧化物催化剂(Oxide Catalyst)影响机制，并关注新的在氧化物催化剂(Oxide Catalyst)新的增长变量。'1133117'项目主要研究的是关于增强碳纤维(Carbon Fiber)/碳纳米管(Carbon NanoTube, CNT)分散颗粒的催化效率的问题。两个项目都是研究碳纳米管在催化剂领域的问题，由于研究目的一致，所以两个项目聚拢在一起。

图 5(b) 中变化比较明显的还有'1046519'和'1434824'两个项目，这两个项目明显聚拢在一起。通过进一步分析发现'1046519'项目研究的是利用高度结晶(Crystaline)的双壁碳纳米管(Double Wall Carbon NanoTube, DWCNT)制备具备高强度和韧性的新一代结构纤维。'1434824'项目研究的是一种新型碳纳米结构(Carbon NanoStructure)工程过程，称为纳米管融合(NanoTube Fusion)。该方法可以创建高性能的碳纤维(Carbon Fiber)，可以应用于航空航天、高功率密度的能量存储和轻质布线等领域。正是由于两个项目研究的都是碳纳米管在材料纤维方面的应用，所以被

聚拢在一起。

为了对这 6 个项目数据相似度进行两两对比，图 6 给出了 6 个项目数据间的层次距离关系。图 6(a)是原始项目文本之间的层次距离关系，图 6(b)为经过语义组块标注后的层次距离关系。可以看出在没有进行语义组块标注之前进行余弦相似度计算时，有两组关系最为紧密的项目分别为'1046519'和'0933141'；'1133117'和'0945004'。



(a) 原始项目数据层聚类结果　(b) 基于语义组块特征的项目数据层聚类结果

图 6　文本相似度两两对比结果

通过深入分析这些项目就会发现:

(1) 第一组中，'1046519'项目主要研究利用高度结晶(Crystaline)的双壁碳纳米管(Double Wall Carbon NanoTube, DWCNT)制备具备高强度和韧性的新一代结构纤维。而'0933141'项目研究开发一种新的纳米晶(Nano-Crystalline)混合金属氧化物催化剂(Oxide Catalyst)，能够获得理想的导电性和电化学特性。这两个项目的研究目的有着根本不同。

(2) 第二组中，'1133117'主要研究的是关于增强碳纤维(Carbon Fiber)/碳纳米管(Carbon NanoTube, CNT)分散颗粒的催化效率的问题。'0945004'项目研究利用凝胶纺丝技术(Gel Spun Technology)开发一种高强度–低密度的碳纳米管(Carbon NanoTube, CNT)基碳纤维(Carbon Fiber)。一个研究催化剂，一个研究碳纤维，两个项目的研究目的也不一样。

分析产生这种现象的原因，本文认为，在利用余弦相似度计算过程中，将文本数据生成向量空间模型(VSM)，在 VSM 中把所有词语的重要程度等同看待，并且忽略各词语的语义特性。由于在 NSF_CNT 数据集中的项目数据都是研究碳纳米管领域的项目，在词汇分布上都会出现诸如"Carbon"，"NanoTube"等共性词语，所以利用传统的余弦相似度计算模型难以将这些项目数据真正的语义相似度计算出来。

在图 6(b)中，可以同样发现两组最为紧密的数据分别为: '1133117'和'1007793'；'1046519'和'1434824'。这与图 5(b)中的结果是一致的。

(1) 第一组中，'1133117'和'1007793'两个项目都是研究催化剂的问题，两个项目的区别只是在研究方法和思路上不同。'1133117'项目主要研究碳纳米管(Carbon NanoTube, CNT)本身做为非均相催化反应(Heterogeneous Catalysis)尤其是在 FT 催化反应中的催化剂(Catalysts)的作用。'1007793'项目主要探寻石墨烯(Graphene)和碳纳米管(Carbon NanoTube, CNT)对氧化物催化剂(Oxide Catalyst)影响机制问题。

(2) 第二组中，'1046519'和'1434824'两个项目都是研究碳纤维的问题，而且应用领域也基本集中在商业航空航天等领域。两个项目追求的强度、密度等技术参数指标也基本一致。两个项目的区别也是在研究方法和思路上不同。'1046519'研究利用高度结晶(Crystaline)的双壁碳纳米管(Double Wall Carbon NanoTube, DWCNT)制备具备高强度和韧性的新一代结构纤维。该纤维可以为车辆防弹、商业航空航天等领域提供强度更高、重量更轻的结构纤维材料。'1434824'研究的重点是一种新型碳纳米结构(Carbon NanoStructure)工程过程，称为纳米管融合(NanoTube Fusion)。该方法可以创建高性能的碳纤维(Carbon

Fiber)，可以应用于航空航天、高功率密度的能量存储和轻质布线等领域。

5 结 语

与原始余弦相似度计算模型相比，本文提出的基于语义组块特征的改进Cosine文本相似度计算方法可以有效提升文本间语义相似度。此外，由于该模型能够区分句子中词汇的语义角色，可以有效消除噪音数据的影响，并且可以降低向量空间模型维度，提升计算效率。与基于本体的计算方法相比，该模型不需要外部的本体库支持，模型的泛化能力也得到提高。有效的文本相似度计算方法可以发现不同文档之间的相互关系，通过计算基金项目数据或者论文文本之间的相似度可以有效地挖掘出文本之间存在的主题关联性，进而可以深入分析识别科技创新过程中的知识扩散过程、新兴研究前沿主题出现等。下一步将开展针对不同语义组块分别进行不同权重的文本相似度计算研究，从不同维度分析文本相似度，实现文本相似度细粒度分析。

参考文献：

[1] Salton G, Wong A, Yang S, A Vector Space Model for Automatic Indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.

[2] 孙建军, 成颖. 信息检索技术[M]. 北京: 科学出版, 2004. (Sun Jianjun, Cheng Ying. Information Retrieval Technology [M]. Beijing: Science Press, 2004.)

[3] Jacob B, Benjamin C. Calculating the Jaccard Similarity Coefficient with Map Reduce for Entity Pairs in Wikipedia [EB/OL]. [2017-04-07]. http://www.infosci.cornell.edu/weblab/papers/Bank2008.pdf.

[4] Metzler D, Bernstein Y, Croft W B, et al. Similarity Measures for Tracking Information Flow[C]// Proceedings of the 14th ACM International Conference on Information and Knowledge Management. 2005:517-524.

[5] Banerjee S, Pedersen T. Extended Gloss Overlaps as a Measure of Semantic Relatedness[C]// Proceedings of the 17th International Joint Conference on Artificial Intelligence. New York: ACM Press, 2003: 805-810.

[6] Ponzetto P S, Strube M. Knowledge Derived from Wikipedia for Computing Semantic Relatedness[J]. Journal of Artificial Intelligence Research, 2007, 30(1): 181-212.

[7] Allan J, Bolivar A, Wade C. Retrieval and Novelty Detection at the Sentence Level[C]// Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. 2003.

[8] Landauer T K, Foltz P W, Laham D. Introduction to Latent Semantic Analysis[J]. Discourse Processes, 1998, 25(2-3): 259-284.

[9] Lund K, Burgess C. Producing High-dimensional Semantic Spaces from Lexical Co-occurrence[J]. Behavior Research Methods Instruments & Computers, 1996, 28(2): 203-208.

[10] Islam A, Inkpen D. Semantic Text Similarity Using Corpus-based Word Similarity and String Similarity[J]. ACM Transactions on Knowledge Discovery from Data, 2008, 2(2): 10.

[11] Sébastien H, David S, Sylvie R, et al. A Framework for Unifying Ontology-based Semantic Similarity Measures: A Study in the Biomedical Domain[J]. Journal of Biomedical Informatics, 2014, 48(2): 38-53.

[12] Rada R, Mili H, Bicknell E, et al. Development and Application of a Metric on Semantic Nets[J]. IEEE Transactions on Systems, Man, and Cybernetics Society, 1989, 19(1): 17–30.

[13] Leacock C, Chodorow M. Combining Local Context and WordNet Similarity for Word Sense Identification[M]. MIT Press, 1998.

[14] Pekar V, Staab S. Taxonomy Learning: Factoring the Structure of a Taxonomy into a Semantic Classification Decision[C]// Proceedings of the 19th International Conference on Computational Linguistics, Taipei, Taiwan, China. New York: ACM Press, 2002: 1-7.

[15] Wu Z, Palmer M. Verb Semantics and Lexical Selection [C]//Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics. New York: ACM Press, 1994: 133-138.

[16] Tversky A. Features of Similarity[J]. Psychological Review, 1977, 84(4): 327-352.

[17] Wang J Z, Du Z, Payattakool R, et al. A New Method to Measure the Semantic Similarity of GO Terms[J]. Bioinformatics, 2007, 23(10): 1274-1281.

[18] Couto F M, Silva M, Coutinho P M. Implementation of a Functional Semantic Similarity Measure Between geNe-products[D]. Lisbon: University of Lisbon, 2003.

[19] Othman R M, Deris S, Illias R M. A Genetic Similarity Algorithm for Searching the Gene Ontology Terms and Annotating Anonymous Protein Sequences[J]. Journal of

Biomed Information, 2008, 41(1): 65-81.
[20] 李文清, 孙新, 张常有, 等. 一种本体概念的语义相似度计算方法[J]. 自动化学报, 2012, 38(2): 229-235. (Li Wenqing, Sun Xin, Zhang Changyou, et al. A Semantic Similarity Calculation Method of Ontology Concept [J]. Acta Automatica Sinica, 2012, 38(2): 229-235.)
[21] 刘宏哲, 须德. 基于本体的语义相似度和相关度计算研究综述[J]. 计算机科学, 2012, 39(2): 8-13. (Liu Hongzhe, Xu De. Review of Semantic Similarity and Correlation Calculation Based on Ontology [J]. Computer Science, 2012, 39(2): 8-13.)
[22] 黄承慧, 印鉴, 侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J]. 计算机学报, 2011, 34(5): 856-864. (Huang Chenghui, In Jian, Hou Fang. A Text Similarity Measure Based on Semantic Information and TF-IDF Method [J]. Journal of Computers, 2011, 34(5): 856-864.)
[23] 刘宏哲. 文本语义相似度计算方法研究[D]. 北京: 北京交通大学, 2012. (Liu Hongzhe. Text Semantic Similarity Calculation Method [D]. Beijing: Beijing Jiaotong University, 2012.)
[24] 白如江. 基于语义计算的科学研究前沿识别研究[D]. 北京: 中国科学院大学, 2015. (Bai Rujiang. Scientific Research Frontier Recognition Research Based on the Semantic Computing [D]. Beijing: Chinese Academy of Sciences, 2015.)

作者贡献声明:

白如江: 设计研究方案, 采集、清洗、分析实验数据;
冷伏海: 提出研究思路;
廖君华: 起草及修改论文。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: brj@sdut.edu.cn。
[1] 白如江, 冷伏海, 廖君华. NSF_CNT.RAR. NSF 碳纳米管项目相似度计算原始数据.
[2] 白如江, 冷伏海, 廖君华. COSINE_similarity.py. 改进 Cosine 相似度计算方法与可视化展示 Python 程序.
[3] 白如江, 冷伏海, 廖君华. NSF_CNT_POS_CHUNK.CVS. NSF 项目数据语义组块标注结果数据.

收稿日期: 2017-04-27
收修改稿日期: 2017-05-17

An Improved Cosine Text Similarity Computing Method Based on Semantic Chunk Feature

Bai Rujiang[1] Leng Fuhai[2] Liao Junhua[1]
[1](Institute of Scientific and Technical Information, Shandong University of Technology, Zibo 255049, China)
[2](Institute of Policy and Management, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: [**Objective**] This paper aims to improve the performance of Cosine text similarity computing method with the help of text semantic chunk feature. [**Methods**] First, we retrieved the project data of carbon nanotubes studies, which were pre-processed with stemming and POS techniques. Then, we identified the semantic chunk of text contents with the conditional random field model. Third, we calculated the similarity of texts based on semantic chunk feature. Finally, we compared our results with those generated by the unlabeled data. [**Results**] The proposed method improved the performance of Cosine similarity calculation by up to 26%. [**Limitations**] Our study relies on semantic chunks to annotate the computing performance. [**Conclusions**] The proposed method could effectively identify similar texts, and reduce the dimensions of vector space model, which improves the computing efficiency. The new method is robust and could be transferred to other fields.

Keywords: Text Similarity Semantic Chunks Vector Space Model Ontology